Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.080178

ARTICLE

Research on Agricultural Machinery Fault Nested Entity Extraction for Low-Resource and High-Noise Scenes

Huaixuan Yan and Yan Gong*

School of Light Industry, Harbin University of Commerce, Harbin, China

*Corresponding Author: Yan Gong. Email: gyan163@163.com

Received: 04 February 2026; Accepted: 13 May 2026; Published: 15 June 2026

ABSTRACT: To correctly diagnose faults in farm machinery, we need to know a lot about the field and have experience with maintenance. However, most of this important information is stored in old, unstructured documents like technical manuals and expert logs. These documents don't have a standard way to be represented digitally, which makes it very hard to build automated diagnosis systems. There are three main technical problems with getting structured knowledge out of this kind of text: noise from optical character recognition (OCR) during digitization, the extreme lack of labeled samples in specialized fields (low-resource constraints), and the complex nested structures that are common in descriptions of mechanical components. To fill this gap in research, this paper suggests a semantic-enhanced nested entity extraction framework that is made for situations with few resources and a lot of noise. To fill this gap in research, this paper suggests a semantic-enhanced nested entity extraction framework engineered specifically for low-resource and high-noise constraints. First, to mitigate the severe visual noise inherent in digitized legacy documents, we introduce a Targeted Noise-Injection Denoising Paradigm. This module utilizes whole-word masking to simulate and correct OCR character confusion prior to feature extraction. Second, to overcome extreme data sparsity, we propose a Dynamic Domain-Constrained Augmentation Algorithm. Governed by a TF-IDF-weighted substitution formula, this algorithm mathematically isolates and preserves high-information domain entities while expanding the syntactic feature space. Finally, we architect a Hierarchical Span-Decoding Network. By integrating contextual word embeddings with bidirectional temporal gating and a global pointer matrix, this network transcends the "flat" assumptions of traditional sequence labeling to accurately identify multi-level nested entities, such as parts-assembly relationships. Experimental results demonstrate that the proposed framework achieves an F1-score of 95.87% with minimal seed data. Ablation studies also show that the data augmentation strategy leads to big performance gains. Moreover, by employing this method, we create a fault knowledge graph comprising 19,710 entities and validate the efficacy of converting unstructured text into computable fault knowledge via a Retrieval-Augmented Generation (RAG) system.

KEYWORDS: Agricultural machinery; knowledge graph; fault diagnosis; named entity recognition; data augmentation; nested entity extraction; retrieval-augmented generation

1 Introduction

As farming machines get smarter and more accurate, it becomes harder to find problems with important parts like the transmission and chassis systems. To do this, you need to use deep semantic mining on huge, unstructured technical documents like maintenance manuals and expert logs. Also, the ongoing improvement of artificial intelligence methods has pushed fault diagnosis toward very complicated, self-driving models. Recent literature underscores the burgeoning potential of spatial-channel collaborative multi-scale graph interaction deep transfer learning [1], adaptive model-agnostic meta-learning networks [2], and

convolutional-transformer reinforcement learning agents [3] to transform predictive maintenance. These advanced, data-driven models are the most advanced way to diagnose problems in industry. To fully use their deep inference abilities in real-world farming situations, it is very important to give them highly structured, domain-specific knowledge as a second step [4]. Therefore, building a domain-specific Knowledge Graph (KG) serves as an essential foundational bridge. By turning noisy, unstructured texts into clean, computable knowledge, we provide the robust data structures necessary to fuel both current automated diagnosis systems and future advanced reinforcement learning agents. In contrast to general or financial domains, the implementation of Named Entity Recognition (NER) in agricultural machinery fault diagnosis faces three distinct technical challenges that current mainstream sequence labeling models (e.g., BiLSTM-CRF) find difficult to resolve.

First, recognizing nested entity structures is a big problem. Agricultural machinery systems have a built-in "assembly-part" hierarchy, where fault descriptions often include nested entities. For example, in "[[hydrostatic variable speed pump] swash plate] wear", the "swash plate" is a part that is part of the "hydrostatic variable speed pump" assembly. The majority of current research depends on Conditional Random Field (CRF)-based [5] sequence labeling architectures, which are limited by the presumption of flat entities. Because of this, these models cannot properly decode overlapping entity boundaries, which means that important hierarchical structural information is lost during extraction.

Second, digitizing legacy documents creates a significant amount of "semantic noise". In this study, we quantitatively define a "high-noise" scenario as processing text corpora that exhibit an Optical Character Recognition (OCR) Character Error Rate (CER) exceeding 15%. This digitization process inevitably causes character confusion, such as unit symbol errors or garbled text, which severely compromises the semantic integrity of professional terminology. Most current models lack the mechanisms to rectify semantic errors caused by this level of noise, making them highly unreliable when working with low-quality industrial text.

Lastly, the field suffers from "Low-Resource" generalization, which we strictly define as scenarios restricted to fewer than 500 manually annotated seed training samples. Due to the scarcity of domain experts [6], real-world engineering environments rarely offer large-scale annotated corpora. Operating within this extreme low-resource constraint, directly applying data-driven deep learning models is highly susceptible to overfitting, meaning the models will fail to generalize well to unseen samples.

To address these intricate, domain-specific challenges, we propose a synergistic semantic-enhanced nested entity extraction framework. Inductive Bias Alignment is the most important new method in this framework. In agriculture, traditional sequence labeling models don't work because their linear inductive bias doesn't fit with the nested, hierarchical reality of mechanical assemblies. Our framework uniquely aligns the model architecture with the physical domain: the upstream denoising module mirrors the visual degradation of the source documents, while the downstream span-based matrix decoder physically mirrors the nested, overlapping topology of agricultural machinery. The primary methodological contributions of this paper are as follows:

- Novel Denoising Paradigm via Targeted Noise-Injection: Rather than relying on generic pre-trained error correction, we designed a custom Domain-Adaptive Pre-training (DAPT) mechanism. By mathematically simulating OCR visual-character confusion distributions, we force the semantic representation layer to develop resilience to the specific degradation patterns of legacy industrial texts.
- Dynamic Domain-Constrained Augmentation Algorithm: We introduce a novel augmentation algorithm governed by a dynamic TF-IDF-weighted substitution formula (p_w). This mathematically isolates and "freezes" high-information domain entities while mutating low-entropy syntax, solving the critical dilemma of semantic drift in few-shot industrial regimes.

- Hierarchical Span-Decoding Network: To align with the "assembly-part" nested structures of the agricultural domain, we designed a span-based decoding architecture. By utilizing whole-word contextual embeddings coupled with bidirectional temporal gates and a global pointer matrix, this network transforms Named Entity Recognition (NER) into a multidimensional spatial prediction task [7], effectively bypassing the theoretical limitations of the Conditional Random Field (CRF) flat assumption.

2 Related Work

Entity recognition methods for fault diagnosis have changed a lot over the years. They started with rule-based systems, then moved on to statistical learning, and most recently to deep learning [8]. There are three main types of these methods: traditional rule-based and statistical models, deep learning sequence labeling models, and pre-training-based approaches.

In the first category, early studies mostly used rule templates and dictionary matching technologies. Domain experts usually made specific fault terminologies and used string matching algorithms to get information out of these methods [9]. These methods work well in closed, canonical situations, but they don't work well in general because they rely too much on manually defined feature rules. This makes it hard to adapt to the different ways people use language in unstructured text. Later, statistical machine learning techniques like Hidden Markov Models (HMM) and Support Vector Machines (SVM) slowly took the place of rule-based systems in industrial text mining. But these models are still limited by the time-consuming process of feature engineering, which makes it hard to process huge and complicated fault descriptions.

Deep learning has made end-to-end sequence labeling models the most common way to do things. The BiLSTM-CRF architecture, introduced by Dong et al. (2016) [10], is regarded as the most exemplary model in this domain. It uses a Bidirectional Long Short-Term Memory network to capture contextual features and Conditional Random Fields (CRF) to limit label transfer, setting a long-lasting standard in the field.

In the "pre-training + fine-tuning" age, researchers have looked into how BERT and its variations might be able to help with fault diagnosis. Gao et al. (2021) suggested a recognition technique [11] that integrates RoBERTa-wwm-ext with deep learning to tackle polysemy in industrial equipment fault data. This research presented the Whole Word Masking strategy to improve the model's semantic representation of Chinese-specific terms, showing better results than traditional BERT models on self-constructed datasets. To further resolve complex features in industrial texts, Zhang et al. (2024) designed an LEBERT-CRF model that integrates a Lexicon Adapter into the Transformer layer [12], effectively alleviating boundary ambiguity in professional compound words. In the specific field of agricultural machinery, Zhao et al. (2022) used the ALBERT lightweight pre-trained model [13] and shared parameters to make the memory footprint much smaller, making it easier to use on embedded diagnostic devices.

Even with these advancements in pre-training, directly applying existing mainstream methods to agricultural machinery fault diagnosis remains highly problematic due to two compounded limitations. The first major limitation is the inability of standard models to resolve complex, overlapping entity boundaries. Most of the aforementioned works (such as Gao et al. and Zhang et al.) rely on CRF decoding, which is fundamentally bound by the "Flat Assumption"—meaning each token can only receive a single, mutually exclusive label. Consequently, these models entirely miss the "assembly-component" nested structures (e.g., "[[hydrostatic drive pump] swash plate] wear") inherently characteristic of agricultural mechanical hierarchies.

Recent research in industrial text mining has started to move away from flat NER and toward nested entity extraction because flat NER has some problems. For example, recent research in aerospace maintenance and railway fault diagnosis has investigated span-based extraction techniques to obtain overlapping

component structures. In the same way, Machine Reading Comprehension (MRC) paradigms and boundary-aware pointer networks have been suggested as ways to clear up multi-level boundary ambiguities in general manufacturing logs.

However, this introduces the second major limitation: these state-of-the-art nested NER models are notoriously data-hungry and assume access to pristine, manually cleaned corpora. They are ill-equipped to handle the severe data degradation present in the agricultural sector. Existing nested models do not account for the high-frequency Optical Character Recognition (OCR) noise generated during the digitization of legacy farming manuals, nor do they possess mechanisms to prevent overfitting in extreme low-resource (few-shot) scenarios.

To fill this specific research gap, this paper presents a semantic-enhanced nested entity extraction framework specifically engineered for the intersection of structural complexity and data degradation. By coupling a GlobalPointer-based nested decoder with an upstream MacBERT semantic correction module and domain-constrained data augmentation [14], we resolve the limitations of both traditional CRF sequence labeling and modern, data-dependent nested extraction networks.

3 Semantic Error Correction and Hierarchical Span-Decoding Architecture

In this section, we detail the architectural design and workflow of the proposed agricultural machinery fault entity extraction framework. Rather than a simple concatenation of models, this framework operates on the principle of Inductive Bias Alignment, utilizing a bottom-up pipeline designed to mitigate the cascading failures typical of unstructured industrial text. The entity extraction framework and workflow are illustrated in Fig. 1.

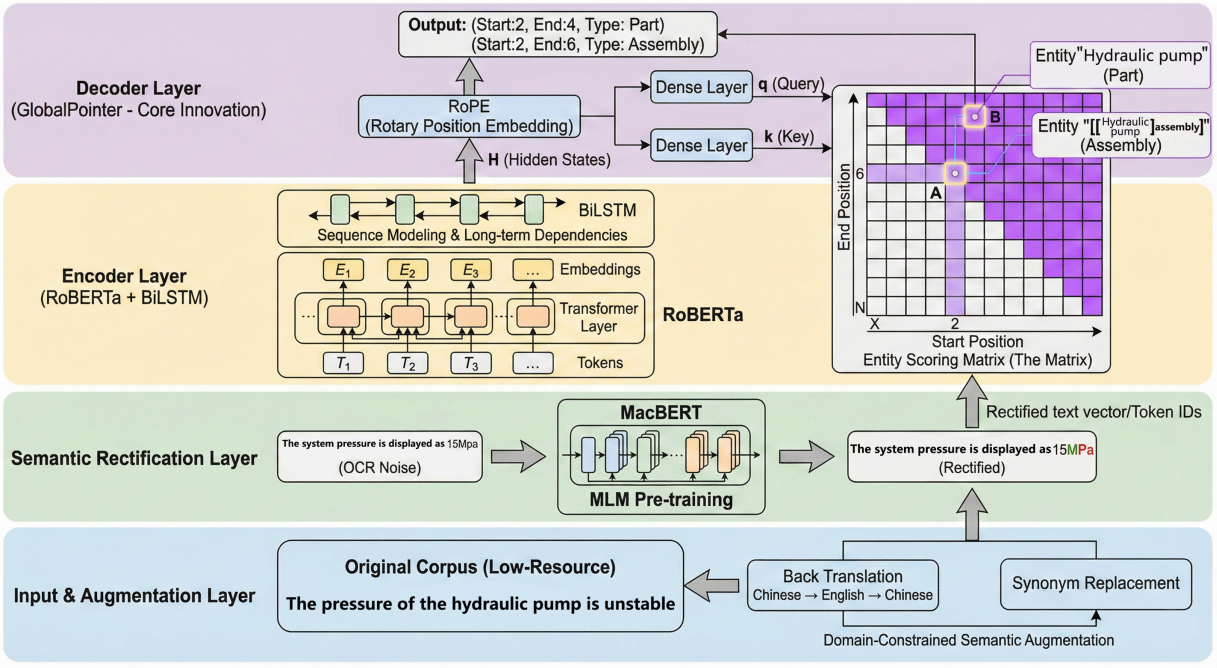

Figure 1: The pipeline framework of the proposed method.

First, we detail the Targeted Noise-Injection Denoising Paradigm (implemented via MacBERT) and the Dynamic Domain-Constrained Data Augmentation strategy [15]. We demonstrate how mathematically

simulating OCR noise creates the clean semantic baseline required to expand the feature space without distorting specialized domain terms.

Second, we analyze the Contextual and Temporal Feature Extraction Layer (implemented via RoBERTa-wwm-ext and BiLSTM). We explain how this layer utilizes dynamic word embeddings and bidirectional temporal modeling to capture the long-range causal dependencies of complex mechanical failures.

Finally, we construct the Span-Based Matrix Decoder (implemented via GlobalPointer). We present the mathematical derivation showing how Rotary Positional Embeddings (RoPE) and a global pointer matrix utilize these pristine contextual features to accurately identify overlapping entity boundaries [16], effectively resolving the hierarchical extraction of nested mechanical structures.

3.1 Semantic Rectification and Data Augmentation Strategies

This study develops a semantic preprocessing module to tackle the dual issues of Optical Character Recognition (OCR) noise and data sparsity in the diagnosis of agricultural machinery faults [17,18]. This module uses adversarial pre-training and limited sample augmentation to give downstream entity extraction tasks high-quality, high-density feature inputs.

To fix character-level confusion that happened when old technical documents were digitized (for example, when "50 MPa" was misidentified as "50 Mpa" or "valve" was misidentified as "va1ve"), we use MacBERT (MLM as Correction BERT) to make a way to fix semantic errors. The regular BERT model uses the special [MASK] token for Masked Language Model (MLM) pre-training. MacBERT, on the other hand, fixes the problem where OCR errors show up as "false substitutions" instead of "deletions". By combining Whole Word Masking (WWM) with N-gram masking and replacing masked tokens with synonyms instead of [MASK] tokens [19], MacBERT improves the masking strategy and makes the noise distribution more realistic.

Let the input sequence be $X = \{x_1, x_2, \cdots, x_n\}$. For a chosen set of mask positions M, the model uses a thesaurus S to replace $x_i (i \in M)$ and make the simulated noisy input $\tilde{X}$. The goal of optimization is to make the conditional probability of the original token $x_i (i \in M)$ given the noisy context $\tilde{X}$ as high as possible. This is how the loss function is defined:

$$\mathcal{L}_{MLM} = -\sum_{i \in M} \log P\left(x_i \mid \tilde{X}; \theta\right) \tag{1}$$

where θ stands for the model's settings. This pre-training task serves as a "denoising autoencoder", allowing the model to effectively identify semantic inconsistencies in context, rectify errors caused by OCR, and preserve the semantic integrity of professional terminology.

After fixing the semantics, we deal with the low-resource problem, which is that we only have 240 seed data points (Few-shot scenario). We create a hybrid data augmentation pipeline that includes "iterative back-translation" and "synonym replacement" to lower the risk of overfitting that comes with training deep neural networks on small datasets.

The Iterative Back-Translation strategy uses English, a language with a lot of resources, as a pivot language to make a "Chinese→English→Chinese" loop. Let s be the original sample, and let M_{fwd} and M_{bwd} be the forward and backward translation models, respectively. The process of generation is made official as follows:

$$s' = M_{bwd}\left(M_{fwd}(s)\right) \tag{2}$$

This process adds syntactic diversity (e.g., voice conversion) while keeping the original semantic logic, which effectively increases the syntactic feature space.

We also use a Domain-Constrained Synonym Substitution strategy to stop semantic drift [20], which is when domain-specific proper nouns are changed (for example, changing "hydrostatic variable speed pump" to "hydrostatic changing speed pump", which breaks the entity boundary). This method is applicable with Non-Entity Tokens and uses TF-IDF weights to figure out the replacement probability on the fly. The probability of replacing a token w, p_w, is defined as being inversely proportional to its TF-IDF value:

$$p_w = \frac{\lambda}{\mathrm{TF-IDF(w)} + \epsilon} \tag{3}$$

where λ is a normalization coefficient and ϵ is a term that smooths things out. As a result, it's common for words that don't add much information to be replaced, while important domain terms are kept. This hybrid strategy effectively increases the size of the training corpus, which makes the model much more generalizable and robust when there aren't many resources available.

3.2 The RoBERTa-wwm-ext Encoding Layer

After rectifying and semantic boosting to the semantics, making a high-quality feature space is very important for turning text into computable representations. Descriptions of problems with agricultural machinery often use long, high-frequency professional compound nouns, such as "hydrostatic variable speed pump" or "combine harvester". Because they can't deal with polysemy, traditional static word vector models (like Word2Vec) aren't good for this job [21]. Also, the basic BERT model deals with polysemy by being aware of context, but its built-in character-level masking often messes up the meaning of technical terms. For example, hiding "harvest" in "combine harvester" might make the model only look at local co-occurrences like "combine" and "machine" and not the bigger picture.

To fix this, we use RoBERTa-wwm-ext as the encoder that both sides use. The main benefit of this model is the Chinese Whole Word Masking (WWM) feature, which makes it so that the whole word "harvester" is hidden as a single semantic unit. This strategy makes the model pay attention to more general information in the context, like "crop lodging" in the text before this one or "threshing cylinder" in the text after this one, in order to figure out what entities are missing.

Assume that the preprocessed input sequence is $X = \{x_1, x_2, \cdots, x_n\}$, where n is the longest sequence length. The RoBERTa model has L stacked layers of bidirectional Transformer encoders. First, the input sequence is mapped to an initial embedding representation H_0, which is made up of the element-wise sum of Token, Position, and Segment embeddings:

$$H_0 = E_{\mathrm{token}} + E_{\mathrm{pos}} + E_{\mathrm{seg}} \tag{4}$$

The Multi-Head Self-Attention (MHSA) mechanism then spreads information hierarchically to capture global dependencies. For layer l ($1 \le l \le L$), the process includes calculating attention and changing the feed-forward network. The input H_{l-1} is transformed into the Query (Q), Key (K), and Value (V) matrices through a linear mapping [22]. Scaled dot-product attention is used to figure out how relevant the context is:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

where $\sqrt{d_k}$ is the factor that scales. The model concatenates outputs from multiple attention heads, followed by Layer Normalization and Residual Connections, in order to capture multi-dimensional semantic features.

$$H'_l = \text{LayerNorm}\left(H_{l-1} + \text{MultiHead}\left(H_{l-1}\right)\right) \tag{6}$$

Lastly, a position-wise Feed-Forward Network (FFN) changes intermediate features H'_l in a non-linear way:

$$H_l = \text{LayerNorm}\left(H'_l + \text{FFN}\left(H'_l\right)\right) \tag{7}$$

After L layers of deep interaction, we obtain the final context semantic representation matrix $H_L \in \mathbb{R}^{n\times d}$, where d = 768. This matrix dynamically fuses context information with literal character meanings, serving as the foundation for the subsequent BiLSTM layer.

3.3 BiLSTM Feature Extraction Layer

The Transformer-based RoBERTa model is great at capturing global semantics, but agricultural machinery technical logs need strong modeling of local sequential features because they have short sentence patterns and strange punctuation [23]. To help the Transformer better capture local sequential dependencies, we add a Bidirectional Long Short-Term Memory (BiLSTM) network on top of the embedding layer. This layer transforms the global contextual features from RoBERTa into dynamic sequence features with temporal dimensions.

There are two separate LSTM units in the BiLSTM: Forward and Backward. If the input vector at time t is x_t (the t-th row of the RoBERTa output H_L), the forward LSTM encodes information h_t from the start of the sequence, and the backward LSTM encodes $\overleftarrow{h}_t$ from the end:

$$h_t = \text{LSTM}_{fwd}\left(x_t, h_{t-1}\right), \overleftarrow{h}_t = \text{LSTM}_{bwd}\left(x_t, \overleftarrow{h}_{t+1}\right) \tag{8}$$

To solve the vanishing gradient problem, each LSTM unit controls the flow of information through a gating mechanism. The internal state update is defined as:

$$\begin{aligned}
f_t &= \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right)\\
i_t &= \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right)\\
\tilde{C}_t &= \tanh\left(W_C \cdot \left[h_{t-1}, x_t\right] + b_C\right)\\
C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t\\
o_t &= \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right)\\
h_t &= o_t \odot \tanh\left(C_t\right)
\end{aligned} \tag{9}$$

The forget, input, and output gates are represented by f_t, i_t, and o_t, respectively. The candidate cell state is $\tilde{C}_t$, the current cell state is C_t, the Sigmoid activation function is σ, and the Hadamard product is $\odot$. The last hidden state, v_t, is the combination of the two-way vectors:

$$v_t = \left[h_t; \overleftarrow{h}_t\right] \tag{10}$$

This "Global-Local" feature fusion helps the model accurately show the causal logic chain between the "fault phenomenon" (the thing that happens) and the "fault cause/solution" (the thing that happens as a result), giving nested entity decoding a lot of valuable context.

3.4 GlobalPointer Decoding Layer

The "Flat Assumption" in traditional Conditional Random Field (CRF) decoding means that each token can only have one label that doesn't overlap with any other label. This presents a theoretical constraint for agricultural machinery fault descriptions, characterized by nested structures (e.g., within the assembly entity "hydrostatic variable speed pump", the component "swash plate" is identified as a nested entity). To address this structural constraint, we put forward the GlobalPointer decoding mechanism, reconfiguring Named Entity Recognition (NER) from sequence labeling to a Span-based Global Matrix Prediction challenge. A comparison between flat NER and nested NER in named entity recognition is shown in Fig. 2.

(1) Interaction of Rotary Position Embedding (RoPE): For each entity category α, GlobalPointer builds an N × N scoring matrix. The value of each element $s_\alpha(i, j)$ tells us how sure we are that the span from position i to j is a valid entity. We use Rotary Position Embedding (RoPE) to clearly show relative position information [24]. Let v_i and v_j be the feature vectors at positions i and j for an entity of type α (the outputs from the BiLSTM layer). We first make query (q) and key (k) vectors by doing a linear transformation:

$$q_{i,\alpha} = W_{q,\alpha} v_i, k_{j,\alpha} = W_{k,\alpha} v_j \tag{11}$$

Then, a rotation matrix R changes these vectors so that the inner product score can be found:

$$s_\alpha(i, j) = \left(\mathcal{R}_i q_{i,\alpha}\right)^T \left(\mathcal{R}_j k_{j,\alpha}\right) \tag{12}$$

One of the main things about RoPE is that $\mathcal{R}_i^T \mathcal{R}_j = \mathcal{R}_{j-i}$. This makes sure that the score only depends on the relative distance (*j*-*i*), which is very vital for recognizing entities because entity validity is based on span length and internal constituents, not absolute coordinates. The chance $p_\alpha(i, j)$ is given by:

$$P_\alpha(i, j) = \text{sigmoid}\left(s_\alpha(i, j)\right) = \frac{1}{1 + e^{-s_\alpha(i,j)}} \tag{13}$$

(2) Matrix Decoding and Multi-Label Classification: We build a prediction space that is upper triangular (because the end position j is greater than or equal to i). This matrix lets us activate more than one overlapping area at the same time. For example, if the coordinates (2, 5) and (2, 8) both go over the threshold, the model finds two nested entities that share a start position. The training goal uses a Circle Loss variant to make the space between target entity spans (S_{pos}) and non-entity spans (S_{neg}) as wide as possible:

$$\mathcal{L} = \log\left(1 + \sum_{(i,j)\in S_{\text{pos}}} e^{-s_\alpha(i,j)}\right) + \log\left(1 + \sum_{(k,l)\in N_\alpha} e^{s_\alpha(k,l)}\right) \tag{14}$$

GlobalPointer can capture all levels of fault components at the same time during a single forward pass thanks to this matrix-based strategy. The complete model architecture is shown in Fig. 3, with the corresponding layer interactions detailed in Table 1.

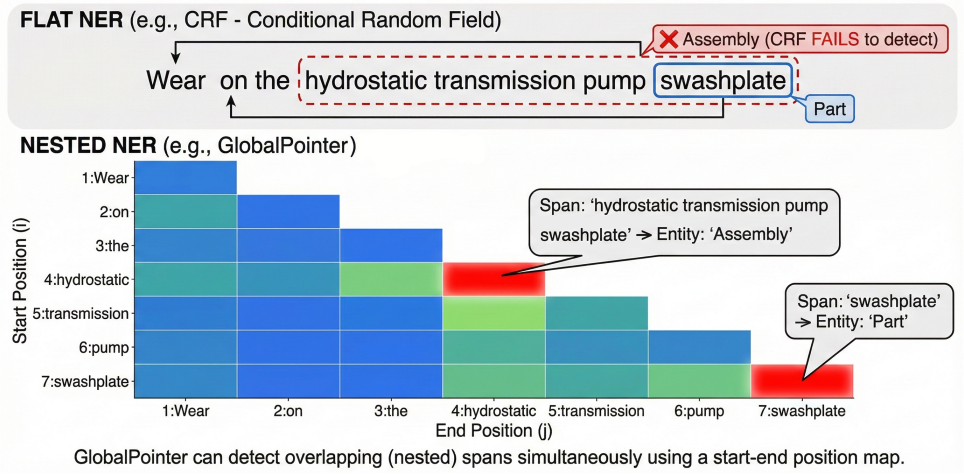

Figure 2: Comparison of boundary detection capabilities.

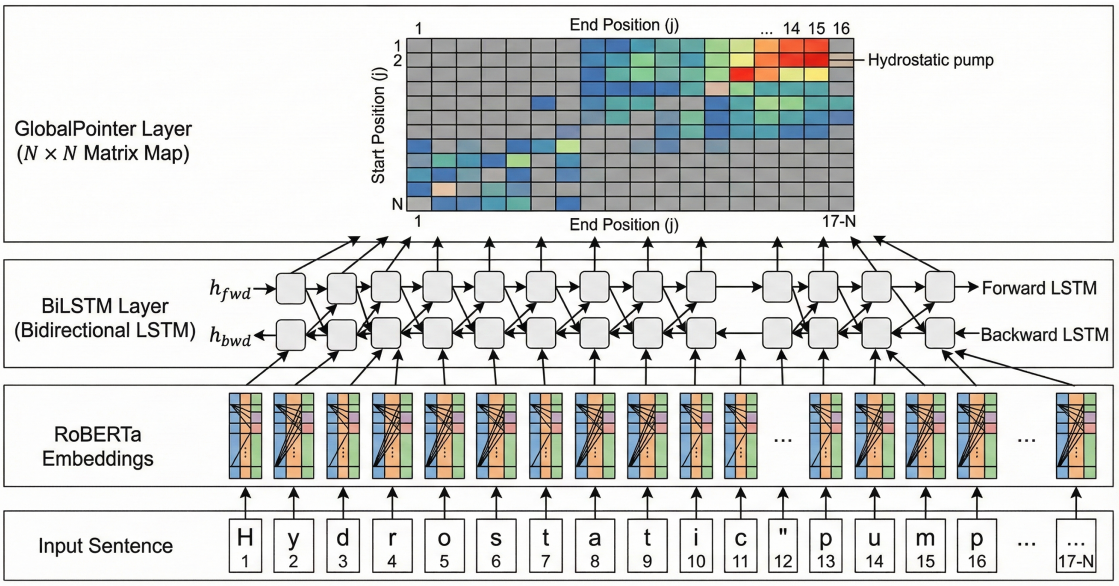

Figure 3: RoBERTa-BiLSTM-GlobalPointer architecture.

Table 1: Comparative functional analysis of the model layers.

Layer	Core Mechanism	Primary Functional Objective
Input Layer	Chinese WWM (Tokenizer)	Captures domain-specific mechanical semantic units
Encoding Layer	RoBERTa-wwm-ext	Generates dynamic, context-aware embeddings
Feature Layer	BiLSTM	Models long-range causal and temporal dependencies
Decoding Layer	GlobalPointer (GP)	Resolves nested/overlapping entity boundary detection

4 Experiments

The goal of this section is to thoroughly test the proposed model's ability to extract agricultural machinery fault entities through a systematic empirical study. We first detail the construction of the domain-specific dataset, including data acquisition channels, ontology definitions, and preprocessing strategies. Next, we describe the experimental setup, which includes the hardware and software configurations, the evaluation protocols, and the baseline models. Finally, we confirm the framework's effectiveness and robustness under low-resource constraints and high-noise interference by doing a lot of comparative experiments and component-level ablation studies.

4.1 Building the Agricultural Machinery Fault Dataset

The generalization performance of deep neural networks is significantly contingent upon high-quality domain-labeled data. This study creates a special fault diagnosis dataset for specialized fields to deal with the lack of open, standardized Chinese fault corpora in the agricultural machinery field.

(1) Data Acquisition and Digitization: For our main data sources, we chose official technical manuals and frontline maintenance logs from well-known companies that make agricultural machinery, like John Deere and Kubota. First, Optical Character Recognition (OCR) technology was used to transcribe unstructured legacy documents. Despite utilizing high-precision OCR tools, the initial transcription yielded a Character Error Rate (CER) of approximately 18%, firmly placing our dataset within the defined "high-noise" threshold (CER > 15%). Following transcription, a heuristic rule-based filter utilizing regular expressions was applied to remove non-textual noise. To resolve the severe data quality degradation associated with this traditional digitization, we integrated the proposed MacBERT semantic error correction module to automatically rectify OCR-induced character confusion.

(2) Defining and annotating entities: We made a strict Fault Ontology Schema based on the knowledge of experts in the field. This schema has four main entity categories: "Fault Component", "Fault Phenomenon", "Fault Cause", and "Solution".

We implemented a "Human-in-the-Loop" active learning strategy to ensure rapid and accurate annotation under strict resource constraints. Domain experts first annotated exactly 240 seed samples at a fine-grained level to serve as a cold-start set. This initial set falls well below our 500-sample quantitative boundary, establishing a rigorous "low-resource" regime for evaluating the framework. Subsequently, a pre-trained model was employed for preliminary labeling, followed by manual verification. The final benchmark dataset comprises 1807 entity samples. Fig. 4 shows how entities are spread out across categories in a statistical way.

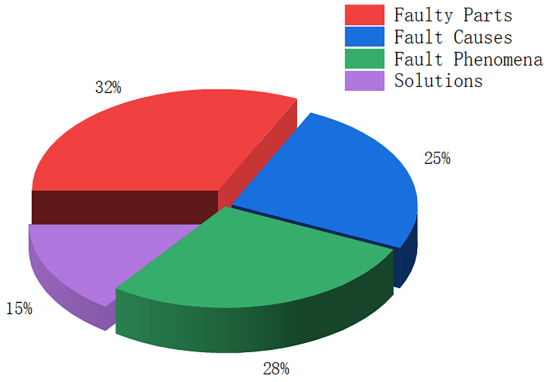

Figure 4: Distribution of entity annotations.

4.2 Putting Data Augmentation Algorithms into Action

To effectively reduce the risk of overfitting due to a lack of seed data, we systematically grew the original small-sample corpus based on the semantic enhancement strategy described in Section 3.1.

(1) We used a neural machine translation system to create a dual-channel translation loop (Chinese → English → Chinese). We added random changes during generation by setting the temperature parameter to $\tau = 0.7$. This helped us get the most syntactic diversity while keeping the meaning consistent. This setting makes nondeterministic sampling easier, which makes the augmented text more expressive.

(2) Synonym Replacement: To keep domain-specific terminology intact, we made a custom dictionary of 1200 domain stop words (like standard part names and failure modes) to make sure that proper noun boundaries weren't broken. This vocabulary acts as a filter to "freeze" core entity tokens. For all other non-entity tokens, the replacement probability p_w is inversely related to the TF-IDF weight. The normalization coefficient λ and a smoothing term ϵ control this relationship.

To prevent division by zero for terms with extremely low inverse document frequencies and to cap the maximum substitution rate, the smoothing term ϵ is fixed at a constant 10^{-5}. The normalization coefficient λ dictates the overall intensity of the augmentation and was rigorously determined through a grid search on the validation set, exploring the parameter space $\lambda \in [0.05, 0.30]$ with a step size of 0.05.

When we were tuning the model, we noticed that values that were too conservative ($\lambda < 0.10$) didn't add enough syntactic variance, which made the model vulnerable to the original overfitting constraints. On the other hand, aggressive values ($\lambda > 0.20$) caused a lot of semantic drift by replacing important context words that define the fault logic. This led to a big drop in the validation F1-score. So, $\lambda = 0.15$ was chosen as the best threshold. This setup makes sure that only general terms with a lot of information and a lot of frequency are replaced, which maximizes corpus diversity while strictly keeping the causal semantic rules of the agricultural machinery domain.

This hybrid augmentation pipeline increased the original training set by about 11 times, bringing the total number of samples in the final experimental corpus to 2160. Table 2 shows that the augmented dataset greatly improves the variety of sentence patterns and vocabulary coverage while keeping the core fault logic intact.

Table 2: Comparative functional analysis of the model layers.

Partition	Source	Sentences	Characters	Entities	Description
Original	Expert annotation	192	6547	854	Few-shot Base
Augmented	Semantic enhancement	2112	72,015	9394	Final Training
Validation	Expert annotation	24	816	108	Hyperparameter Tuning
Testing	Expert annotation	24	829	112	Performance Evaluation

4.3 Metrics for Evaluation

We utilize Precision (P), Recall (R), and F1-score (F1) as the main metrics to fully evaluate how well the model works. Because nested entity structures are so common, we use a strict Exact Match criterion: a prediction is only correct if the entity boundary and category label match the ground truth exactly. The metrics are defined like this:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{15}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{16}$$

$$F1 = \frac{2 \times P \times R}{P + R} \times 100\% \tag{17}$$

where TP (True Positive) means correctly predicted entities, FP (False Positive) means false predictions, and FN (False Negative) means missed entities. We also use the Relative Improvement Rate (RIR) to measure how much better the semantic data augmentation strategy makes things:

$$RIR = \frac{M_{\text{augmented}} - M_{\text{baseline}}}{M_{\text{baseline}}} \times 100\% \tag{18}$$

where M stands for the specific evaluation metric.

4.4 Metrics for Evaluation

To ensure fairness, comparability, and reproducibility, all experiments—including baseline models and ablation variants—were conducted within a strictly unified software and hardware environment.

(1) Hardware and Software Environment: All training and inference tasks were performed on a high-performance computing cluster equipped with a single NVIDIA GeForce RTX 3090 GPU (24 GB VRAM) to accommodate the fine-tuning of large-scale pre-trained models. The software environment operates on Windows 11, utilizing PyTorch 1.12.0 as the deep learning framework and Hugging Face Transformers 4.28.0 for loading pre-trained weights. To eliminate randomness in small-sample experiments, the global random seed was fixed at 42, ensuring deterministic initialization, data loading, and dropout masking.

(2) Core Hyperparameter Settings: We determined the globally optimal parameter combination via a Grid Search strategy based on validation set performance. We selected Chinese-RoBERTa-wwm-ext (12 layers, 12 heads, hidden dimension 768) as the pre-training base. The maximum sequence length was set to 256 with truncation and zero-padding. We employed the AdamW optimizer with a Layer-wise Learning Rate strategy: the Encoder layer (RoBERTa) uses a lower rate (2×10^{-5}) to prevent catastrophic forgetting, while the Decoder layer (BiLSTM & GP) uses a higher rate (5×10^{-5}) to accelerate domain adaptation. Key hyperparameters are listed in Table 3.

(3) Domain-Adaptive Pre-training (DAPT) of Semantic Rectifier: Prior to the downstream extraction task, we conducted DAPT on the MacBERT model using 500 MB of unlabeled technical text (standard documents, maintenance Q&A). To simulate real OCR error distributions, we constructed a Noise-Injection Masking Mechanism. Beyond random masking (15%), this mechanism explicitly replaces tokens with visually similar typos from a Confusion Set with a 10% probability. The model was trained for 50 epochs (Batch Size = 64) and subsequently frozen as a static error correction module, significantly reducing the Character Error Rate (CER) of the raw corpus.

(4) Training Strategy: Given the "Few-shot" nature of the data, we implemented a composite optimization strategy. We implemented linear warmup, where the learning rate increases linearly from zero during the first 10% of steps to stabilize early-stage exploration of the loss surface. Additionally, we utilized an early stopping mechanism (Patience = 3) based on validation set F1 scores, automatically restoring the best checkpoint to balance fitting ability and generalization.

Table 3: Comparative functional analysis of the model layers.

Parameter	Value	Description
Batch Size	16	Adapts to memory constraints and ensures gradient stability
Epochs	30	Total training rounds
Learning Rate (Encoder)	2e–5	Learning rate for the RoBERTa backbone
Learning Rate (Decoder)	5e–5	Learning rate for BiLSTM & GP layers
Dropout Rate	0.1	Prevents overfitting in fully connected layers
Max Gradient Norm	1.0	Gradient clipping threshold to prevent gradient explosion
Positional Embedding Dim	64	Dimension of Rotary Positional Embeddings (RoPE)

4.5 Metrics for Evaluation Assessment of the Efficacy of Data Augmentation Strategies

We did a series of controlled experiments to quantitatively test how well the proposed Domain-Constrained Semantic Enhancement Strategy worked to reduce overfitting in small-sample regimes and make models more robust. The experiments rigorously preserved uniformity in model architecture (RoBERTa-BiLSTM-GP) and hyperparameter settings, designating the size and distribution of the training dataset as the independent variable. The Control Group (Baseline) received training solely on the original small-sample dataset, whereas the Experimental Group (Ours) was trained on the augmented dataset subsequent to semantic enhancement. To guarantee objectivity and practical relevance, both models were assessed on an independent, unaugmented, expert-annotated test set. This protocol seeks to validate the generalization efficacy of the augmentation strategy through performance comparison on novel, real-world samples.

Quantitative Analysis: The results of the evaluation, shown in Fig. 5 and Table 4, show that there are big differences in performance in the low-resource scenario.

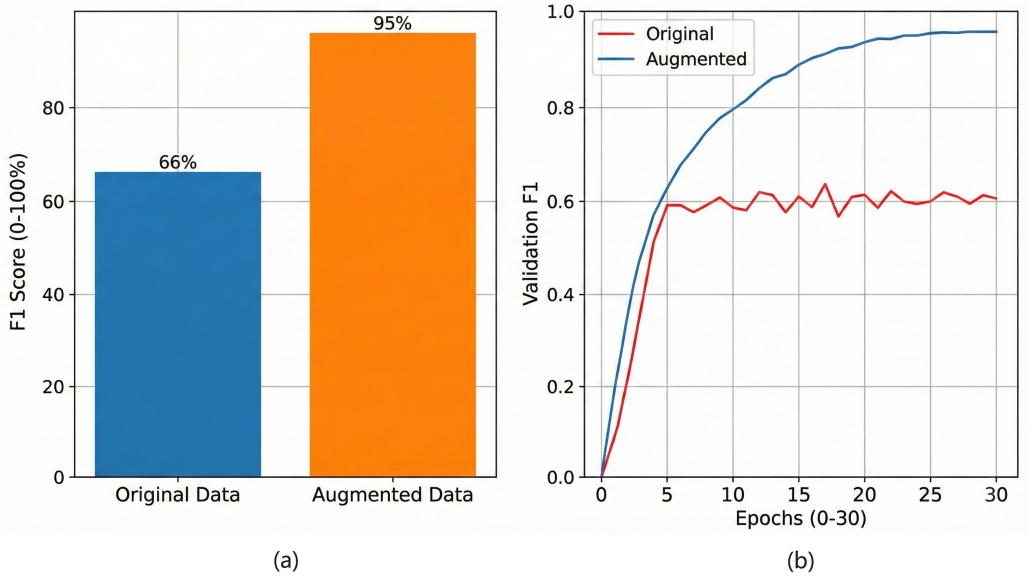

Figure 5: (**a**) Comparison of F1-scores before and after data augmentation; (**b**) comparison of training curves on the validation set.

Table 4: Comparison of model performance before and after semantic data augmentation.

Configuration	Precision	Recall	F1-Score	RIR
Baseline	63.60%	69.58%	66.46%	–
Ours	95.36%	96.39%	95.87%	+44.25%

The model has serious problems with generalization without augmentation. Even though RoBERTa has a strong pre-training base, the baseline model's F1-score stays at 66.46%, and its precision is as low as 63.60%. This result shows that deep models have a hard time making clear decision boundaries in feature spaces that are very sparse. They tend to make a lot of false positives.

On the other hand, the Semantic Augmentation Strategy makes all evaluation metrics much better. The F1-score goes up to 95.87%, which is a Relative Improvement Rate (RIR) of 44.25%. Precision goes up by almost 32 percentage points (from 63.60% to 95.36%), and recall goes up by about 27 percentage points (from 69.58% to 96.39%). These big improvements show that the augmentation strategy works to cover more of the feature space, which helps the model learn more useful features for telling different entities apart.

To learn more about how data augmentation affects optimization dynamics, we looked at the training loss descent curves and validation F1-score trajectories, which are shown in Fig. 6.

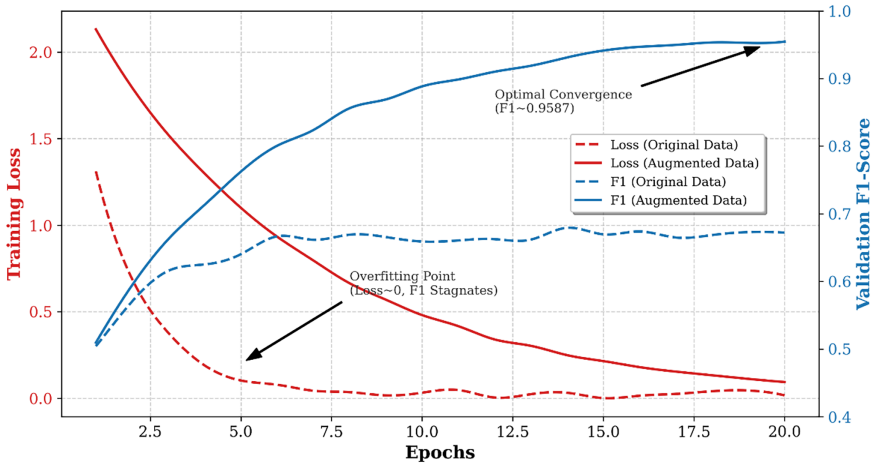

Figure 6: Comparison of training dynamics.

A comparative analysis shows two different ways that things come together:

(1) Memorization Overfitting (Baseline): When training on original data (dotted line), the training loss drops quickly, getting close to zero by the fifth epoch. The validation F1-score, on the other hand, does not improve in the same way; instead, it stays the same or goes up and down. This difference between "high fit on the training set" and "low performance on the validation set" is a sign of typical memorization overfitting. The model memorizes specific high-frequency tokens (like specific bearing models) but doesn't learn abstract semantic representations of "faulty components".

(2) Robust Generalization (Ours): On the other hand, when we train with more data (solid line), the training loss goes down more slowly, but the validation F1-score keeps going up until it reaches a high value. This shows that the different augmented samples do a good job of smoothing out the non-convex loss landscape. This regularization effect stops the model from relying too much on certain entity

vocabulary and makes it focus on strong contextual syntactic features, like causal connectives like "due to". This makes generalization much better in complicated, real-world situations.

4.6 Gradual Optimization of Neural Network Architectures

Based on the real-world evidence presented in the previous sections, the Domain-Constrained Semantic Augmentation Strategy has successfully gotten past the data sparsity problem, creating a strong base for training deep learning models. This section, therefore, focuses on optimizing the architecture to improve the accuracy of fault entity extraction and make it easier to build a high-quality knowledge graph.

We strictly follow a controlled experimental design to make sure that the evaluation is fair and accurate. In this section, all architectural comparative experiments are consistently fine-tuned on the "Semantically Augmented Training Set" and tested on the independent "Expert-Labeled Test Set". This experimental design removes variability caused by data size, which lets us test how well different Pre-trained Language Models (PLMs) encode domain terminology and how well different decoding strategies handle complex nested structures. This empirical analysis reveals RoBERTa-BiLSTM-GP as the superior architecture for this task.

4.6.1 Comparison of Semantic Representation Layers: The Benefit of Whole Word Masking

In the specialized field of diagnosing faults in agricultural machinery, target entities often show up as long-span compound nouns or phrases with complicated attribute constraints, like "hydrostatic CVT drive axle" or "high-pressure common rail injector". This puts a lot of pressure on the pre-trained model's ability to represent meaning. To ascertain the efficacy of various pre-training mechanisms, we executed a comparative analysis of three mainstream Chinese PLMs—Standard BERT-Base, ERNIE (Knowledge Graph enhanced), and RoBERTa-wwm-ext—while maintaining a constant downstream decoder (BiLSTM-GP).

Quantitative Analysis: Quantitative Analysis: Fig. 7 and Table 5 show the results in detail. They show that RoBERTa-wwm-ext has the best overall performance, with an F1-score of 93.80%. This is a lot better than both standard BERT-Base and ERNIE.

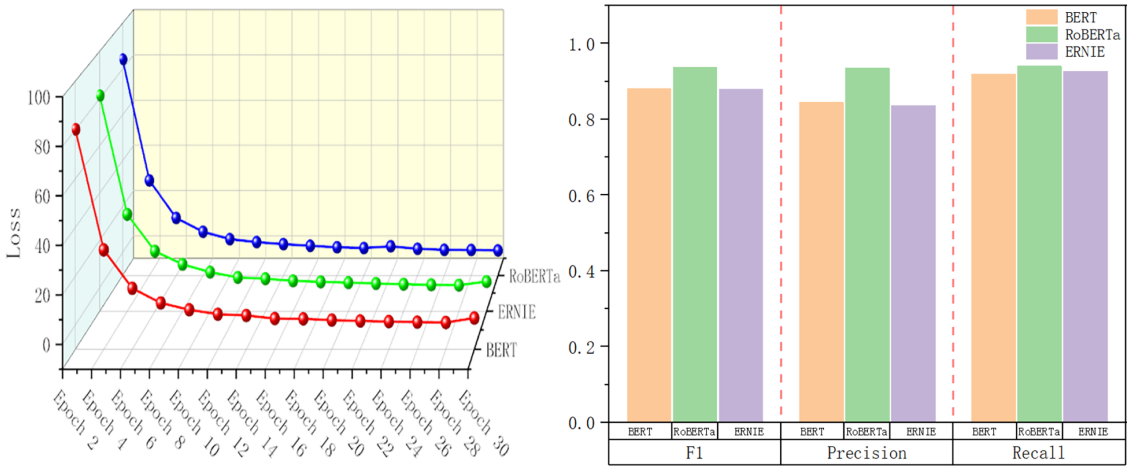

Figure 7: (**a**) Comparison of loss curves across representation layers; (**b**) comparison of performance metrics across representation layers.

Table 5: Performance comparison for the representation layer.

PLM	Masking Strategy	Precision	Recall	F1-Score
BERT-Base	Character-level	0.8448	0.9198	0.8807
ERNIE	Entity-level	0.8367	0.9271	0.8796
RoBERTa-wwm-ext	Whole Word	0.9355	0.9406	0.9380

Mechanism Analysis: The cause of this difference in performance is the alignment between the masking strategy and the structure of the domain terms.

(1) Standard BERT: Uses character-level masking to randomly hide Chinese characters. This strategy often breaks strong internal semantic associations when used with agricultural terms (for example, by splitting "oil pump" into separate "oil" and "pump" tokens). This makes the model learn local character co-occurrence probabilities instead of entity-level semantic representations.

(2) RoBERTa-wwm-ext: This adds the Whole Word Masking (WWM) feature. This is a strict rule: if you choose to mask a token t_i that belongs to a certain semantic unit W (like "clutch"), you must also mask all of the tokens that make up W at the same time. The following limits the mask probability distribution P_{mask} :

$$P_{\text{mask}}\left(t_j \mid t_i \in W\right) = 1, \quad \forall t_j \in W \tag{19}$$

With this strategy, the model has to go beyond local cues and use more general contextual information (like "power interruption" in the text above) to figure out what the missing phrase is. RoBERTa-wwm-ext keeps the semantic integrity of the domain by clearly separating terms that are morphologically similar but functionally different (for example, "oil pump" vs. "oil sump").

Also, ERNIE doesn't do as well because of domain shift; its pre-training corpus (general Baidu Baike) is very different from the specialized agricultural machinery domain. RoBERTa-wwm-ext, on the other hand, can transfer better between domains because it was trained on a lot of large Chinese corpora.

4.6.2 Performance Assessment of Feature Extraction Layer

After determining that RoBERTa-wwm-ext is the best embedding base, the next important step is to choose a feature extraction layer that can capture long-range temporal dependencies. PLMs use Transformer structures by default, but agricultural machinery fault logs don't always have standard punctuation and have a lot of causal dependencies (for example, "failure X... leads to... result Y"). A specialized temporal modeling layer is essential for reliable entity boundary recognition.

We compared five feature extraction architectures: BiLSTM, BiGRU, IDCNN, Transformer Encoder, and Multi-Head Attention. To ensure a rigorous control variable methodology, all feature extraction baselines in this experiment were uniformly configured with the RoBERTa-wwm-ext encoder upstream and a Conditional Random Field (CRF) decoder downstream.

Quantitative Analysis: The BiLSTM architecture outperforms all other architectures on all metrics, with an F1-score of 94.83%. See Fig. 8 and Table 6 for more information. BiGRU has a slight edge in speed of inference, but its F1-score is slightly lower.

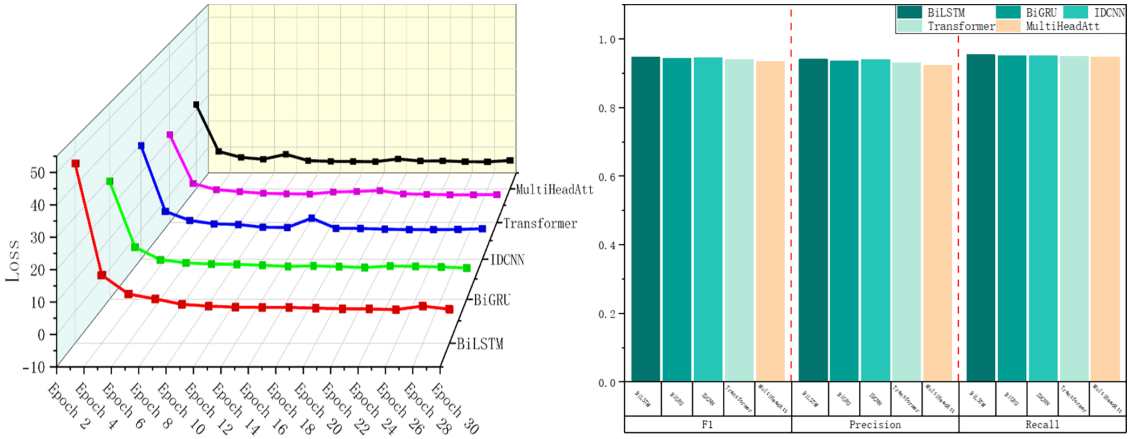

Figure 8: (**a**) Comparison of loss curves across feature extraction layers; (**b**) comparison of performance metrics across feature extraction layers.

Table 6: Performance comparison of feature extraction layers.

Feature Extraction Model	Precision	Recall	F1-Score
BiLSTM	0.9418	0.9549	0.9483
BiGRU	0.9368	0.9519	0.9443
IDCNN	0.9409	0.9519	0.9464
Transformer	0.9314	0.9502	0.9407
MultiHeadAtt	0.9236	0.9488	0.9360

Mechanism Analysis: The internal memory cell structure of BiLSTM makes it better than other models. The state update mechanism is officially defined as:

$$
\begin{aligned}
f_t &= \sigma\left(W_f \cdot \left[h_{t-1}, x_t + b_f\right]\right) \\
C_t &= f_t \odot C_{t-1} + i_t \odot \tanh\left(W_C \cdot \left[h_{t-1}, x_t\right] + b_C\right) \\
h_t &= o_t \odot \tanh\left(C_t\right)
\end{aligned} \tag{20}
$$

Where C_{t-1} is the state of long-term memory and f_t is the gate that lets us forget things. This gating mechanism lets the model choose what information to keep and what to throw away. It solves the vanishing gradient problem in long sequences (200+ tokens) and keeps deep causal links between distant entities.

Notably, models that only use attention (like Transformer and MultiHeadAtt) did poorly. This shows the Inductive Bias Trade-off: Transformers work best with a lot of data, but in situations where there isn't much data (even with augmentation), they can fit noise because they don't make any assumptions about the position of the previous sequence. BiLSTMs, on the other hand, have a structural prior for sequential processing, which lets them generalize syntactic rules more quickly from small samples. In the same way, IDCNN's discrete convolution kernels had a hard time capturing global structural dependencies in nested entities that spanned a long time.

4.6.3 Performance Evaluation of the Decoding Layer: Resolving Nested Entities

The decoding layer is the last part of the process, linking time-based features to specific entity boundaries. Descriptions of agricultural machinery faults often have complicated nested structures. For example, in "[[reaping table] lifting cylinder]", the part "reaping table" is inside the assembly "reaping table lifting cylinder". This is a big problem for standard decoders. We systematically compared Softmax, CRF, and the suggested GlobalPointer. To explicitly isolate the impact of the decoding strategy and ensure a fair baseline comparison, all three decoding modules were universally paired with the optimal upstream configuration determined in the preceding experiments: the RoBERTa-wwm-ext encoder and the BiLSTM feature extraction layer.

Quantitative Analysis: The information in Fig. 9 and Table 7 shows that GlobalPointer works better than CRF and Softmax, with an F1-score of 95.89%, which is 1.28% better than CRF and 5.49% better than Softmax.

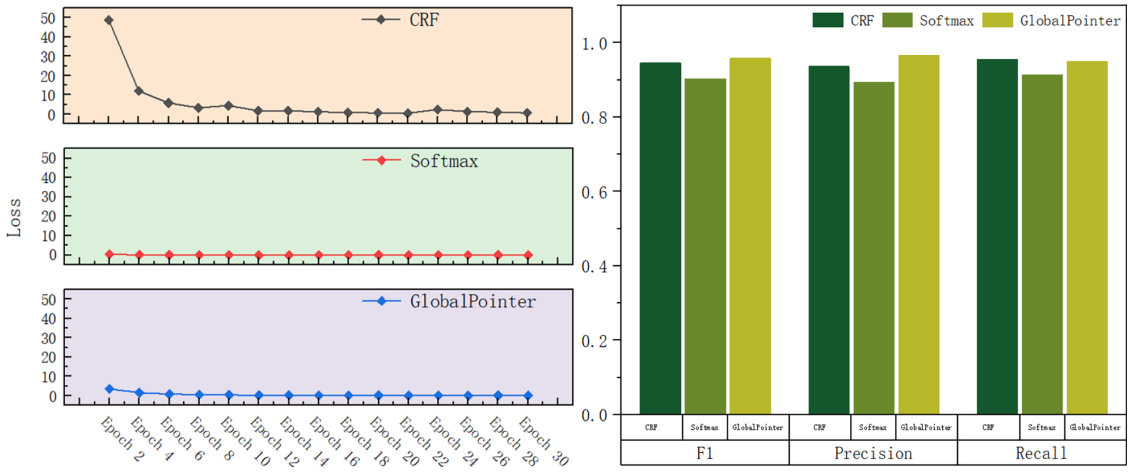

Figure 9: (**a**) Comparison of loss curves across decoding layers; (**b**) comparison of performance metrics across decoding layers.

Table 7: Performance comparison for the decoding layer.

Decoding Strategy	Paradigm	Precision	Recall	F1-Score	Nested Entity Handling
Softmax	Pointwise	89.36%	91.46%	90.4%	Weak
CRF	Sequence	93.67%	95.56%	94.61%	None
GlobalPointer	Span-based	96.7%	95.08%	95.89%	Strong

Mechanism Analysis: We look at the inductive bias of each decoder when it comes to nested structures to explain why this is happening:

(1) Softmax (Pointwise): It assumes that tokens are not related to each other. It can't model label dependency, which leads to a lot of illegal label sequences and the worst performance.

(2) CRF (Sequence-based): CRF is good at modeling label transitions with a state transition matrix A, but it has a big problem when it comes to this task: the "Flat Sequence Assumption". It limits each token x_i to one unique label y_i. Because of this, CRF has to choose between the inner part and the outer assembly for nested terms like "harvester lift cylinder". This causes a lot of false negatives for nested entities.

(3) GlobalPointer (Span-based): This is a big change from sequence labeling to Span-based Matrix Prediction. Instead of putting tokens into groups, it makes an $N \times N$ scoring matrix where the coordinate

(i, j) shows the chance that span $i \rightarrow j$ makes an entity. This design lets you activate overlapping areas at the same time (for example, both (3, 5) and (3, 9)). A targeted evaluation on a nested subset confirmed this benefit: CRF recall dropped to 68.4%, while GlobalPointer stayed at 94.2%. This shows that GlobalPointer's matrix-based mechanism can naturally handle hierarchical structures, which solves the problem of boundary ambiguity.

4.7 Error Analysis and Limitations

The RoBERTa-BiLSTM-GP architecture, enhanced by the semantic data augmentation strategy, attains an impressive F1-score of 95.89% on the test set; however, a residual error rate of approximately 4.1% remains. We did a fine-grained manual audit and attribution analysis on test samples that showed "prediction deviation" to objectively look into the performance limits and find possible failure mechanisms in the current framework. We systematically group the main reasons for failure into three main categories based on the statistical patterns of these errors: optical character confusion, implicit coreference, and the OOV/Long-tail challenge. Fig. 10 shows how different types of errors are spread out, and Table 8 looks at some examples.

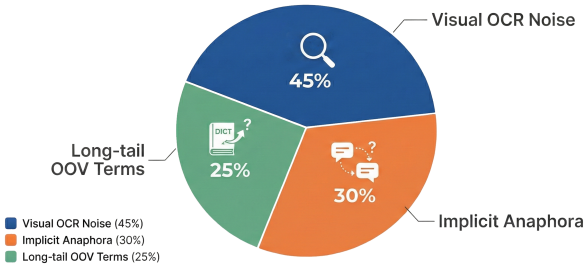

Figure 10: Distribution of error types.

(1) Optical Character Confusion: This type of mistake happens the most often, especially when old paper maintenance logs are turned into digital files. The upstream MacBERT module fixes a lot of logical problems, but there is still a problem with finding morphologically similar characters. In Case 1, presented in Table 8, the term "filter element" was misread as the meaningless term "fl1ter e1ement" because the handwriting was not clear. The RoBERTa tokenizer uses character-based segmentation, so this change at the character level messes up the semantic space of the word embedding. This makes it impossible for the model to map the term to a "faulty component" entity. These kinds of mistakes are basically a loss of multimodal information that text-only models have a hard time fully recovering. This indicates the imperative for subsequent investigations into integrated end-to-end OCR-NER modeling that includes visual attributes.

(2) Cross-sentence Dependency and Implicit Coreference: Missed detections due to cross-sentence dependencies and implicit coreference resolution are also important. A lot of agricultural machinery troubleshooting documents use short engineering language, which often leaves out subjects and uses pronoun anaphora. As shown in Case 2, presented in Table 8, the main assembly that was mentioned earlier (like "oil cylinder") is often left out of the next sentences, which instead talk about sub-components (like "sealing ring"). BiLSTM has the ability to remember things over time, but when the contextual span is longer than the effective memory window or when reference relationships are unclear, the model often misclassifies isolated sub-parts as common nouns instead of specific fault entities. This shows that current sentence-level extraction models have trouble with document-level logic. This means that document-level Graph Neural Networks (GNNs) are needed to improve coreference resolution.

(3) Problems with Out-of-Vocabulary (OOV): Finally, the Out-of-Vocabulary (OOV) problem is a bottleneck that makes it hard to use in real life. Our semantic enhancement strategy adds words to the vocabulary by replacing synonyms, but it can't cover all of the non-standard abbreviations or quickly changing model codes used at maintenance sites. In Case 3, presented in Table 8, the model is careful when it comes to zero-shot terms like "PCV valve" (which are rare English abbreviations) that aren't in the training set. It classifies them as non-entities. This phenomenon illustrates the intrinsic limitation of closed-set training, as the model finds it challenging to adjust to new ideas in an open-world context. To tackle this issue, subsequent research will investigate the incorporation of Retrieval-Augmented Generation (RAG) or Prompt Learning methodologies to equip the model with continuous learning abilities in few-shot or zero-shot scenarios.

Table 8: Performance comparison for the representation layer.

Error Type	Input Text	Prediction	Ground Truth	Failure Mechanism
Visual Noise	Check if the fuel fl1ter e1ement is clogged.	[O] (Non-entity)	[Filter Element] (Part)	Glyph corruption distorts semantic embedding.
Implicit Ref.	...Cylinder leakage. Its sealing ring is aging...	[O] (Non-entity)	[Sealing Ring] (Part)	Failure to resolve anaphora ("its" → "cylinder").
OOV Terms	Replace the PCV valve.	[O] (Non-entity)	[PCV Valve] (Part)	Rare English abbreviation absent from the training set.

5 System Implementation and Validation

The high-precision entity extraction and knowledge graph construction methods described in the preceding sections have established a robust, structured knowledge foundation for agricultural machinery fault diagnosis [25]. To validate the practical utility of this extracted knowledge in real-world engineering scenarios, this chapter details the implementation of a Retrieval-Augmented Generation (RAG) intelligent diagnosis system. This system integrates the Neo4j graph database with a Large Language Model (LLM). This chapter focuses on a comparative analysis of real-world fault cases to demonstrate the critical role of domain knowledge graphs in mitigating hallucinations in large models and enhancing the interpretability of diagnostic results.

5.1 Enhanced Architecture for Knowledge Graph Storage and Retrieval

We used the upstream framework to extract 19,710 entities and 32,279 relational triplets, which we then batch-imported into the Neo4j graph database to make a special knowledge base called Agri-FaultKG. The "Fault Component" is the main hub of the graph topology. It connects "Fault Phenomenon", "Fault Cause", and "Solution" to create a logically coherent causal semantic network.

We used this graph to create a unified "Knowledge Retrieval—Reasoning Generation" architecture. Its workflow is shown in Fig. 11 and consists of three main steps:

(1) Intent Parsing and Entity Linking: When a user sends a natural language query to the system (for example, "How to fix gear engagement problems for the harvester"), the system first uses the RoBERTa-BiLSTM-GP model trained in this study to recognize named entities and correctly parse the main fault elements.

(2) Subgraph Retrieval: After that, the extracted entities are turned into Cypher query statements so that Neo4j can find multiple neighbors and find the subgraph structure that best fits the query intent.

(3) Knowledge Injection and Generation: Finally, the structured triplets that were found are turned into natural language descriptions and added to the ChatGLM context window as "Factual Constraints". This helps the model make diagnostic suggestions that are based on real-world examples and are at an expert level.

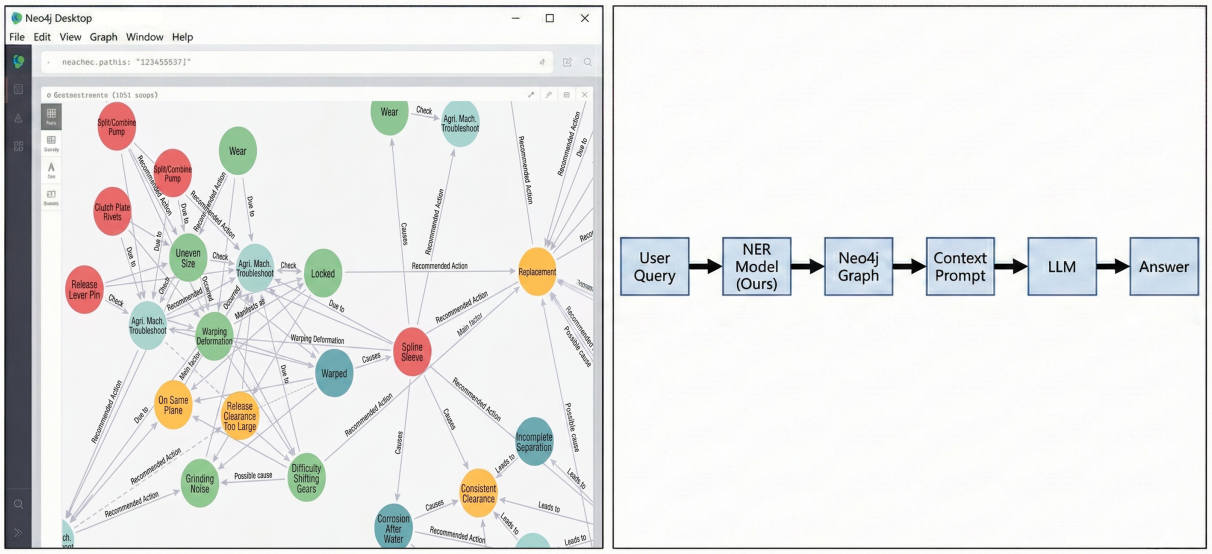

Figure 11: System architecture with graph visualization.

5.2 Comparison of Diagnostic Examples and Effectiveness Analysis

We chose a typical compound fault scenario in agricultural machinery, "the hydraulic system of the combine harvester exhibits high temperature and sluggish propulsion", as a test case to qualitatively assess the system's actual diagnostic performance. We entered this fault description into both a general large model baseline (Standard ChatGLM) and the graph-augmented system we built in this paper (Ours: KG-RAG System). Fig. 12 shows the results of the comparison.

Discussion and Analysis: The comparative results show that the proposed method has a number of clear benefits:

(1) Effective Mitigation of Hallucinations: The general LLM often gives vague, probabilistic answers like "insufficient tire pressure" or "excessive engine load" because it doesn't have enough specialized knowledge in the area. These answers seem to be fluent at first glance, but they don't have a direct causal connection to "hydraulic high temperature", which leads to logical gaps and factual mistakes. On the other hand, the suggested system uses Knowledge Grounding based on clear triplets found in the graph (for example, <sluggish propulsion, cause, wear of oil distribution plate>). It successfully finds the source of the problem in the core assembly, which is called the "Hydrostatic System (HST)". This stops irrelevant responses or hallucinations from happening.

(2) Parameter-Level Precision: Our system gives parameter-level suggestions with engineering reference value, like "adjust the system pressure to 2.5 MPa", instead of the general "check and repair" instructions that most general models give. This is possible because Agri-FaultKG stores precise attribute values.

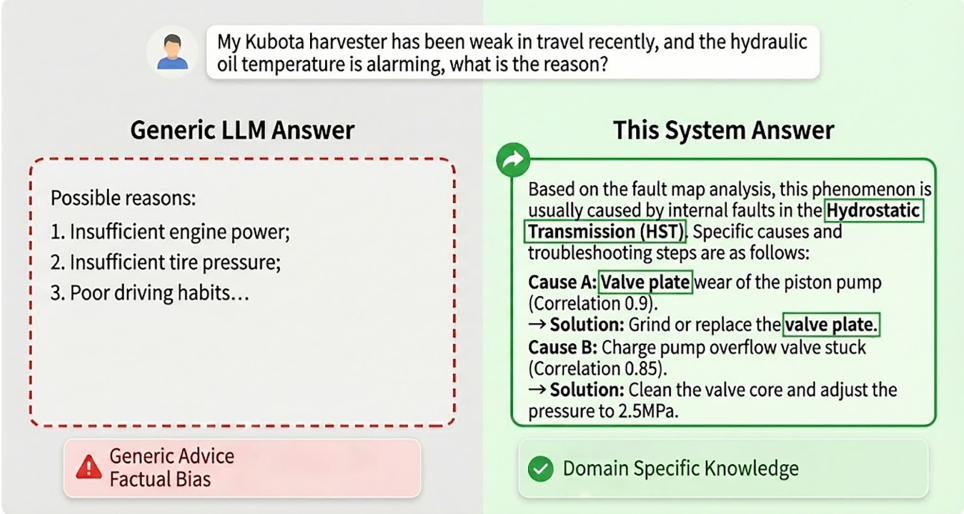

Figure 12: Comparison experiment of diagnostic effect.

In short, the system validation shows that the entity extraction and graph construction methods proposed in this paper are more than just theoretical algorithmic improvements. They turn old maintenance manuals that aren't structured into computable, actionable intelligent decision support. This feature makes it much easier for frontline technicians to troubleshoot complex faults.

6 Conclusion

6.1 Overview of the Research and its Limitations

This study proposed and validated a semantic-enhanced nested entity extraction framework to address three main technical challenges in diagnosing agricultural machinery faults: extreme data sparsity, the noise from optical character recognition (OCR), and the extraction of complex nested entity hierarchies. Instead of using separate extraction tools, this research introduces a unified methodology based on Inductive Bias Alignment.

First, we employed a Targeted Noise-Injection Denoising Paradigm to address character-level ambiguities present in digitized historical documents, thereby creating a clear semantic foundation before proceeding with feature extraction. Second, we formulated a Dynamic Domain-Constrained Augmentation Algorithm governed by TF-IDF-weighted probabilities. This successfully expanded the syntactic feature space while strictly freezing and preserving specialized mechanical terminology to prevent semantic drift in few-shot regimes. Finally, we engineered a Hierarchical Span-Decoding Network. By transforming Named Entity Recognition from a linear sequence labeling task into a global pointer matrix prediction problem, this architecture natively aligns with the physical topology of "assembly-part" relationships, entirely bypassing the theoretical flatness constraints of traditional Conditional Random Fields (CRF).

Empirical assessments reveal that the suggested framework attains an F1-score of 95.89% on the expanded corpus, thereby surpassing established benchmarks. Employing this approach, we constructed Agri-FaultKG, a specialized knowledge graph for the agricultural domain, encompassing more than 19,000

entities. We further validated the practical utility of this computable knowledge by deploying it within a Retrieval-Augmented Generation (RAG) system. Real-world diagnostic testing confirms that grounding large language models in this structured, domain-specific topology effectively mitigates generative hallucinations, proving the viability of transforming highly degraded, unstructured maintenance logs into reliable intelligent decision support.

Even with these significant improvements in low-resource and high-noise environments, our fine-grained error analysis reveals objective boundaries to the current framework's multimodal perception and dynamic adaptability:

(1) The Visual-Semantic Gap: The current denoising paradigm relies exclusively on textual context for error correction. For legacy documents exhibiting extreme visual degradation—such as severe blurring or oil-stained handwriting—a purely text-based model lacks the visual features necessary to reconstruct the true meaning. This "missing modality" represents a physical bottleneck in noise resilience.
(2) Limits of Cross-Domain Generalization: The existing dataset primarily encompasses conventional, broadly applicable agricultural models, such as tractors and typical harvesters. The incorporation of specialized machinery or innovative unmanned equipment, characterized by substantially different structural designs, could potentially trigger a domain shift, thereby affecting the extraction of terminological and fault logic. Consequently, the cross-domain generalization capabilities of the span-based network necessitate additional examination.
(3) Static Knowledge Representation: The present version of Agri-FaultKG serves as a static repository of past knowledge. It does not incorporate a dynamic evolution mechanism capable of autonomously integrating new concepts or removing outdated parameters from real-time maintenance logs. Consequently, the system's capacity to adapt seamlessly within the context of rapidly changing equipment ecosystems is hindered.

6.2 What the Future Holds for Work

To overcome the shortcomings identified in the error analysis and to enhance the technology for the intelligent diagnosis of agricultural machinery, subsequent research will concentrate on the following four principal areas:

(1) End-to-End Multimodal Collaboration: To close the visual-semantic gap that happens when OCR quality drops too low, we plan to use Vision Transformers (ViT) to get features directly from document images. We will explore a "vision-text" dual-stream interaction mechanism to concurrently train Optical Character Recognition (OCR) and Named Entity Recognition (NER), leveraging visual features to facilitate semantic recovery in environments with significant noise.
(2) Document-Level Coreference Resolution: Future architectures must go beyond localized sequence modeling to get around the problems with sentence-level extraction when dealing with implicit coreference and cross-sentence dependencies. We want to add models that can handle longer context windows, like the Longformer. We will also look into how to build a discourse-level reference resolution module that uses Graph Neural Networks (GNNs) to keep track of and resolve anaphora across all maintenance logs.
(3) Dynamic Knowledge Evolution for OOV Terms: To deal with the Out-of-Vocabulary (OOV) problem caused by quickly changing machinery models and non-standard frontline abbreviations, we will build on our current Retrieval-Augmented Generation (RAG) implementation. Our goal is to create a dynamic prompt-learning system that can ask updated manufacturer databases and live maintenance logs questions in real time. This will give the system the ability to learn without any training and keep adding to its knowledge base without having to retrain the model all the time.

(4) Lightweight Modeling for Edge Deployment: Because agricultural machinery vehicle terminals possess strictly limited computing resources, we will investigate lightweight schemes based on knowledge distillation and model quantization. The objective is to significantly reduce the parameter count and inference latency of our models while preserving extraction accuracy, thereby enabling offline, on-board intelligent diagnosis on edge devices.

Acknowledgement: None.

Funding Statement: This research was funded by the Philosophy and Social Sciences Research Planning Project of Heilongjiang Province (23YSD245).

Author Contributions: Study conception and design: Huaixuan Yan, Yan Gong; data collection: Huaixuan Yan; analysis and interpretation of results: Huaixuan Yan, Yan Gong; draft manuscript preparation: Huaixuan Yan, Yan Gong. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The dataset generated and analysed during this study is self-constructed and is not publicly available due to proprietary and commercial restrictions. The data are available on request from the authors.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang X, Jiang H, Dong Y, Mu M. Spatial-channel collaborative multi-scale graph interaction deep transfer learning for unsupervised rotating machinery fault diagnosis. Eng Appl Artif Intell. 2026;176:114691. doi:10.1016/j.engappai.2026.114691.
2. Mu M, Jiang H, Wang X, Dong Y. Adaptive model-agnostic meta-learning network for cross-machine fault diagnosis with limited samples. Eng Appl Artif Intell. 2025;141:109748. doi:10.1016/j.engappai.2024.109748.
3. Li Z, Jiang H, Dong Y. A convolutional-transformer reinforcement learning agent for rotating machinery fault diagnosis. Expert Syst Appl. 2025;271:126669. doi:10.1016/j.eswa.2025.126669.
4. Bai X, Chen Q, Song X, Hong W. Advancing agricultural machinery maintenance: deep learning-enabled motor fault diagnosis. IEEE Access. 2025;13:129933–51. doi:10.1109/ACCESS.2025.3591279.
5. Chen Y, Zhou M, Zhang M, Zha M. Knowledge-graph-driven fault diagnosis methods for intelligent production lines. Sensors. 2025;25(13):3912. doi:10.3390/s25133912.
6. Siddique MF, Zaman W, Umar M, Kim J-Y, Kim J-M. A hybrid deep learning framework for fault diagnosis in milling machines. Sensors. 2025;25(18):5866. doi:10.3390/s25185866.
7. Su J, Murtadha A, Pan S, Hou J, Sun J, Huang W, et al. Global pointer: novel efficient span-based approach for named entity recognition. arXiv:2208.03054. 2022. doi:10.48550/arXiv.2208.03054.
8. Wang C, Sun Y, Wang X. Image deep learning in fault diagnosis of mechanical equipment. J Intell Manuf. 2024;35(6):2475–515. doi:10.1007/s10845-023-02176-3.
9. Thomas G, Balocco S, Mann D, Simundsson A, Khorasani N. Intelligent agricultural machinery using deep learning. IEEE Instrum Meas Mag. 2021;24(2):93–100. doi:10.1109/MIM.2021.9400957.
10. Dong C, Zhang J, Zong C, Hattori M, Di H. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition. In: International Conference on Computer Processing of Oriental Languages. Cham, Switzerland: Springer International Publishing; 2016. p. 239–50. doi:10.1007/978-3-319-50496-4_20.
11. Gao F, Zhang L, Wang W, Zhang B, Liu W, Zhang J, et al. Named entity recognition for equipment fault diagnosis based on RoBERTa-wwm-ext and deep learning integration. Electronics. 2024;13(19):3935. doi:10.3390/electronics13193935.

12. Zhang Z, Yang X, Sun L, Sun Y, Kang J. Research on the knowledge graph for autonomous navigation ship collision accidents based on the enhanced Bert model [Internet]. 2024 [cited 2026 Jan 1]. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4978775.
13. Zhao P, Wang W, Liu H, Han M. Recognition of the agricultural named entities with multifeature fusion based on ALBERT. IEEE Access. 2022;10:98936–43. doi:10.1109/ACCESS.2022.3206017.
14. Li W, Liu J, Gao Y, Zhang X, Gu J. Research on Chinese nested entity recognition based on IDCNNLR and GlobalPointer. Appl Syst Innov. 2024;7(1):8. doi:10.3390/asi7010008.
15. Tang S, Yuan S, Zhu Y. Data preprocessing techniques in convolutional neural network based on fault diagnosis towards rotating machinery. IEEE Access. 2020;8:149487–96. doi:10.1109/ACCESS.2020.3012182.
16. Tang R, Chen Y, Qin Y, Huang R, Zheng Q. Boundary regression model for joint entity and relation extraction. Expert Syst Appl. 2023;229:120441. doi:10.1016/j.eswa.2023.120441.
17. Raza S, Farooq M, Farooq U, Karamti H, Khurshaid T, Ashraf I. A convolutional neural network based optical character recognition for purely handwritten characters and digits. Comput Mater Contin. 2025;84(2):3149. doi:10.32604/cmc.2025.063255.
18. Zhang M, Wang B, Fei H, Zhang M. In-context learning for few-shot nested named entity recognition. In: 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ, USA: IEEE; 2024. p. 10026–30. doi:10.1109/ICASSP48485.2024.10446653.
19. Cui Y, Che W, Liu T, Qin B, Yang Z. Pre-training with whole word masking for Chinese bert. IEEE/ACM Trans Audio Speech Lang Process. 2021;29:3504–14. doi:10.1109/TASLP.2021.3124365.
20. Xu Y, Li S, Feng K, Huang R, Sun B, Yang X, et al. Domain constrained cascadic multireceptive learning networks for machine health monitoring in complex manufacturing systems. J Manuf Syst. 2025;80:563–77. doi:10.1016/j.jmsy.2025.03.021.
21. Fadaee M, Bisazza A, Monz C. Data augmentation for low-resource neural machine translation. arXiv:1705.00440. 2017. doi:10.48550/arXiv.1705.00440.
22. Lou P, Yu D, Jiang X, Hu J, Zeng Y, Fan C. Knowledge graph construction based on a joint model for equipment maintenance. Mathematics. 2023;11(17):3748. doi:10.3390/math11173748.
23. Zhang Y, Zhu Y, Zhu Z, Liu P, Xie P, Wu C. A domain-finetuned semantic matching framework based on dynamic masking and contrastive learning for specialized text retrieval. Electronics. 2025;14(24):4882. doi:10.3390/electronics14244882.
24. Su J, Ahmed M, Lu Y, Pan S, Bo W, Liu Y. Roformer: enhanced transformer with rotary position embedding. Neurocomputing. 2024;568:127063. doi:10.1016/j.neucom.2023.127063.
25. Huang Y, Xue RO, Xu S, Tao HA, Qi SO. Construction and application of a knowledge graph for agricultural motor fault diagnosis. Trans Chinese Soc Agric Eng. 2025;41(6):216–26. doi:10.11975/j.issn.1002-6819.202409065.